RESEARCH ARTICLE | APRIL 24 2024

An approach for modeling and restructuring datasets concerning pollutant emissions in the air FREE

Delyana Dimova

AIP Conf. Proc. 3078, 040007 (2024)
https://doi.org/10.1063/5.0208402

An Approach for Modeling and Restructuring Datasets Concerning Pollutant Emissions in the Air

Delyana Dimova

Department of Mathematics and Informatics, Agricultural University - Plovdiv, 12 Mendeleev Blvd, Plovdiv 4000, Bulgaria

Corresponding author: delyanadimova@abv.bg

Abstract. The examined pollution sources in the current paper include combustion and industrial processes; agriculture; household heating; waste and waste water treatment; road, rail and aviation transport. The investigated group of contaminants released from each of the listed pollution sources for the period 2011-2020 in Bulgaria are the following: sulphur and nitrogen oxides, non-methane volatile organic compounds, methane, carbon and dinitrogen oxide, carbon dioxide and ammonia. A hierarchical organization of the objects from the mentioned sets is built. This studied information from time series stored in xlsx files is structured in a relational database. In this connection, the values of contaminants from a certain pollution source in the relevant file are presented in matrix form. The variation of the number of columns in the formed rectangular matrix depends on the number of the examined years. The data are regrouped to obtain the transposed matrix. The components from each column of this matrix are entered into a selected field of a certain table from the database. The presented approach integrates the processes of searching the indicated data, restructuring and grouping separate datasets, and entering them in the built database. Certain criteria are defined in order to form and extract groups of elements from one or more tables of the database, as well as to establish the relevant dependencies for the studied objects. Analysis of variance and Tukey's range test are used for statistical assessment of the data. The approach can also be applied to examine datasets on air pollutant emissions for other countries.

INTRODUCTION

Information and communication technologies [1], [2] have entered every area of business, economics, management, etc. Most organizations need to collect large amounts of data [3] from electronic and paper sources every year [4]. Different methods are used for their processing, interpretation of the results [5], [6] and making adequate decisions [7], [8] for the considered problems in the respective field.

According to Sharma et al. [9] “Data mining is used to extract implicit and previously unknown information from data”. The study of Priya and Anandhan [10] notes that “Data mining is the process of discovering unrevealed patterns from the huge amounts of data available in database”. In paper [11], the authors present a literature review “on the use of data mining in the analysis of air pollutant measurements”. In addition, some studies consider models for predicting air pollution [12], [13]. Part of these issues are a subject of consideration in the current paper.

The information related to the pollution sources is published on the web site of the Bulgarian National Statistical Institute [14]. These studied data on air pollutant emissions from the considered time series are saved in xlsx files. There are cases where it may take a lot of time for individual users to search certain characteristics for these objects. Existing relationships between objects of examined groups may also be missed. One solution to these problems, in order to avoid the listed above difficulties, is to organize the information from the mentioned time series in one electronic source. In this case, it is appropriate to use a relational database.

The aim of this article is to present an approach for modeling and restructuring datasets concerning air pollutant emissions from major pollution sources in the period 2011-2020 in Bulgaria. Subsequently, the method of the analysis of variance and Tukey's range test are applied for statistical assessment of the studied data. The obtained results are summarized and the relevant conclusions are presented.

12th International Scientific Conference “TechSys 2023” – Engineering, Technologies and Systems
AIP Conf. Proc. 3078, 040007-1–040007-7; https://doi.org/10.1063/5.0208402

MATERIALS AND METHODS

Approach description: The presented hierarchical structure of the elements from the considered datasets is visualized in the scheme from Fig. 1. The study covers ten years time period in Bulgaria. The group of the pollution sources is shown at the first level in the hierarchy. It is denoted as the set S. The mentioned set S = {pollution source_1, pollution source_2, …, pollution source_n} contains n elements, where n>1. The next level involves the pollutants. It is denoted as the set L. The indicated set L = {element_1, element_2, …, element_ j} consists of j components (j>1). It should be noted that each of the investigated sources emits the same pollutants (Fig. 1). The set $C = C_{1j} \cup C_{2j} \cup \ldots \cup C_{nj}$ contains the subsets of the values of the relevant contaminants from each of the considered pollution sources during the studied time period. The components of this set are presented at the next level in the mentioned scheme.

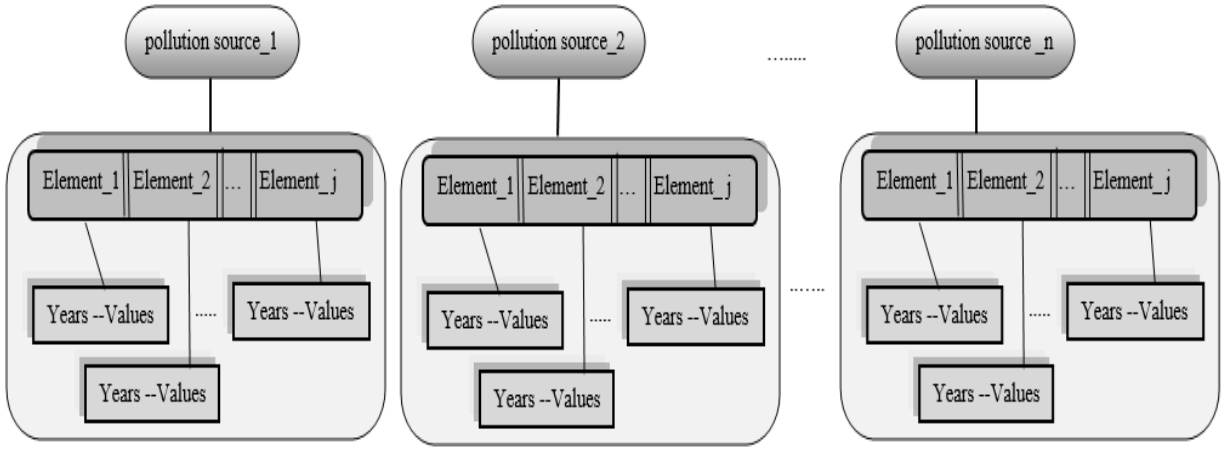

FIGURE 1. General scheme of data organization and hierarchy

In general, the presentation of the data related to a considered pollution source in the mentioned xlsx file is as follows:

$$\begin{pmatrix} c_{11} & c_{12} & c_{13} & \ldots & c_{1p} \\ c_{21} & c_{22} & c_{23} & \ldots & c_{2p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ c_{j1} & c_{j2} & c_{j3} & \ldots & c_{jp} \end{pmatrix} \quad (1)$$

where the formed rectangular matrix contains p rows and j columns, c_{jp} - the value of the j-th pollutant in the p-th year, 1<=p<=10. The variation of the number of columns in the relevant formed matrix depends on the number of examined years.

These studied data need to be restructured. They must be presented in an appropriate form, as they will subsequently be organized and distributed in the relevant tables of a built relational database. For this purpose, the transposed matrix is formed and the following representation is obtained:

$$\begin{pmatrix} c_{11} & c_{21} & \ldots & c_{j1} \\ c_{12} & c_{22} & \ldots & c_{j2} \\ c_{13} & c_{23} & \ldots & c_{j3} \\ \ldots & \ldots & \ldots & \ldots \\ c_{1p} & c_{2p} & \ldots & c_{jp} \end{pmatrix} \quad (2)$$

Practically, the components of each of the columns of this considered matrix are entered into a relevant field of a designed table from the database. In addition, the time segment must be entered. In this connection, the following rectangular matrix is formed:

$$\begin{pmatrix} y_1 & c_{j1} \\ y_2 & c_{j2} \\ y_3 & c_{j3} \\ \dots & \dots \\ y_p & c_{jp} \end{pmatrix} \quad (3)$$

where, y_p - given year from time period, 1<=p<=10.

The current study considers a set from seven pollution sources including:

- combustion processes;
- industrial processes;
- household heating;
- road transport;
- rail and aviation transport;
- waste and waste water treatment;
- agriculture.

The studied set of pollutants released from each of these listed sources are the following eight:

- sulphur oxides;
- nitrogen oxides;
- non-methane volatile organic compounds;
- methane;
- carbon oxide;
- carbon dioxide;
- dinitrogen oxide;
- ammonia.

The information on these considered pollutant emissions during the indicated ten years period from 2011 to 2020 in Bulgaria is organized in the database.

Moreover, it should be noted that the presented approach integrates the processes of searching the data from xlsx files, restructuring and grouping separate datasets, and entering them into the relevant tables of the built database.

The created database includes the following table schemes:

- Continents (ID_continent, Continent);
- Countries (ID_country, Country, ID_continent);
- Pollution sources (ID sources, pollution sources, ID_country);
- Pollutants (ID, Emissions of pollutants, name, ID sources);
- Quantities (ID1, Year, Value (tonne), ID).

In addition, the relationships between the relevant tables are of type one-to-many.

Certain criteria are defined in order to form and extract groups of elements from one or more tables of the database, as well as to establish the relevant dependencies and tendencies for the studied objects. New subsets can contain n selected pollution sources (n>=1) that release the respective pollutant. Separately, other groups are formed from j considered pollutants (j>=1) and related to them data for one or more pollution sources. The difference between the values of each pollutant for the current and previous year is calculated:

$$w_{jp-1} = c_{jp} - c_{jp-1}, \text{ where } 2<=p<=10 \quad (4)$$

The set of values of these p-1 elements (w_{jp-1}) are checked to see if they are only positive or negative. In this connection, the processes regarding the increase or decrease of the considered indicators are investigated, as well as the obtained dependencies and tendencies. The presented results for the indicated data are discussed and summarized. The method of analysis of variance [15], [16] and Tukey's range test [17] are applied in the current study to perform the statistical assessment [18], [19], [20] of these examined data.

RESULTS AND DISCUTION

The result of the organization and distribution of the examined datasets in the built database is visualized in Fig.2. Structuring this examined information in the mentioned database enables users to obtain a quick access to individual elements, as well as a faster search of relevant components. A significant advantage of the built database is the possibility of its expansion. In this regard, the individual levels of structuring of the respective sets of elements do not change, but at the same time new subsets of objects can be added to each of these levels.

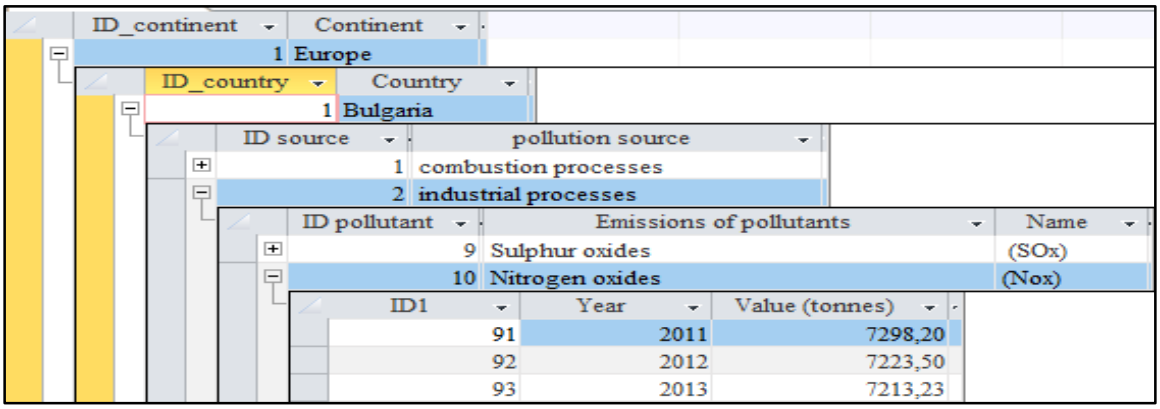

FIGURE 2. Visualization of the structured dataset in the database tables
Source: Data from the National Statistical Institute, Bulgaria

The present work investigated eight contaminants. They are released from each of the considered seven pollution sources. The obtained fifty-six combinations of elements containing a certain contaminant and the relevant pollution source are studied in the period from 2011 to 2020. The calculations show that the subset of the examined differences (w_{jp-1}) between the values of sulphur oxides in the air from combustion processes for the current and previous year are negative for the whole time segment. Analogous results are presented for the considered data regarding ammonia in the air from waste and waste water treatment. The situation is approximately the same with the studied information on elements such as methane and non-methane volatile organic compounds in the air from waste and waste water treatment, as well as nitrogen oxides from combustion processes. But here the calculated differences (w_{jp-1}) for each of the listed three pollutants have positive values for one of the investigated years from the period. For the first and second element, this year is 2012, while for the third one, the year is 2014. The results for the studied data showed a decreasing tendency in the values of each of these five elements during the indicated period. The obtained linear regression models related to these five sequentially listed above contaminants released from the mentioned pollution sources are the following:

$$y = -4652.314x + 93920328.517 \quad (5)$$

$$y = -121.06x + 245769 \quad (6)$$

$$y = -3672.273x + 7518714.397 \quad (7)$$

$$y = -54.009x + 110447 \quad (8)$$

$$y = -6079.795x + 12289353.786 \quad (9)$$

In addition, the constructed models are adequate at 5% level of significance. The calculated decline in the values of sulphur oxides and nitrogen oxides in the air from combustion processes is quite large. It is about 95.25% for the first group and about 81.66% for the second one. The same dependencies are established for the emission values of ammonia, non-methane volatile organic compounds and methane in the air from waste and waste water treatment. But here the decline is about 46.47%, 26.56% and 24.34% for the listed groups, respectively.

The performed summaries show that the examined variables (w_{jp-1}) are positive for several studied datasets during the period 2011-2016. In the case, the considered contaminants are nitrogen oxides, dinitrogen oxide and ammonia in the air from agriculture. The emissions of these pollutants increased by about 63.89%, 51.25% and 17.51% respectively for the listed years. The reverse process is observed for the next group of elements, where the

values of w_{jp-1} are negative in the segment 2016-2020. This group includes carbon oxide and non-methane volatile organic compounds in the air from road transport. Here, a gradual decrease in the values of the listed elements is calculated by about 37.47% and 43.14%. It should be noted that the values of the rest contaminants released from the relevant pollution sources decrease and then increase or vice versa in certain sub-intervals of the indicated time period. Quite naturally, from an environmental point of view, efforts are aimed at reducing air pollutants.

The current study applies the analysis of variance to the investigated datasets related to each pollutant. The results show the presence of statistically proven differences between the emission values of the relevant contaminant released from the considered pollution sources (p<alpha, alpha=0.05). In addition, the formed groups obtained from the application of Tukey's range test are visualized in Table 1.

TABLE 1. The results from the presented evaluations of the studied datasets

Pollution source	Assessment of the emissions of methane			Pollution source	Assessment of the emissions of non-methane volatile organic compounds		
road transport	0,699	a	1	rail and avio transport	334.311	a	1
rail and avio transport	2.968	a		combustion processes	629.502	a	
combustion processes	1743.072	a		waste and waste water treatment	1592.920	a	
household heating	11408.998	b	2	road transport	10602.331	b	2
industrial processes	47789.067	c	3	agriculture	14022.864	c	3
agriculture	78493.080	d	4	household heating	20569.551	d	4
waste and waste water treatment	117248.876	e	5	industrial processes	27159.509	e	5
Pollution source	**Assessment of the emissions of ammonia**			**Pollution source**	**Assessment of the emissions of carbon oxide**		
rail and avio transport	0.099	a	1	agriculture	25.448	a	1
combustion processes	155.887	a		waste and waste water treatment	91.893	a	
road transport	814.4561	ab	2	combustion processes	4937.383	a	
industrial processes	1407.958	bc		rail and avio transport	17111.204	b	2
waste and waste water treatment	1764.841	bc		industrial processes	22343.960	b	
household heating	2160.648	c	3	road transport	64454.905	c	3
agriculture	35474.602	d	4	household heating	154736.499	d	4
Pollution source	**Assessment of the emissions of carbon dioxide**			**Pollution source**	**Assessment of the emissions of nitrogen oxides**		
road transport	7401.296	a	1	waste and waste water treatment	9.908	a	1
waste and waste water treatment	14777.841	a		rail and avio transport	797.794	a	
agriculture	32656.448	a		household heating	2656.533	a	
rail and avio transport	64387.490	a		industrial processes	7250.757	ab	2
household heating	826952.410	a		agriculture	15534.109	b	
industrial processes	4611051.911	b	2	combustion processes	35527.386	c	3
combustion processes	30994325.874	c	3	road transport	40261.294	c	
Pollution source	**Assessment of the emissions of sulphur oxides**			**Pollution source**	**Assessment of the emissions of dinitrogen oxide**		
agriculture	0.190	a	1	road transport	0.1998	a	1
waste and waste water treatment	1.320	a		rail and avio transport	17.679	a	
rail and avio transport	14.350	a		household heating	136.037	a	
road transport	39.690	a		industrial processes	464.451	a	
household heating	6185.0327	a		waste and waste water treatment	478.655	a	
industrial processes	34739.696	ab	2	combustion processes	622.204	a	
combustion processes	146541.736	b		agriculture	13157.156	b	2

Means with the same letter are not significantly different

Source: Own calculations on the basis of data from National Statistical Institute

The presented results from the evaluation of these studied data are the following:

- five groups with statistically proven differences for the considered seven pollution sources are established according to the values of methane released from them. The situation is analogous for the examined data related to the non-methane volatile organic compounds;
- four groups with statistically significant differences for the listed pollution sources are formed depending on the values of the ammonia released from them. A similar case is also obtained for the investigated emissions of carbon oxide;
- three groups are presented with proven differences for the indicated sources releasing carbon dioxide into the air. The number of obtained groups is the same for the examined data on nitrogen oxides;
- two groups with significant differences for the mentioned pollution sources are formed according to the values of sulphur oxides released from them. The case is similar for the studied data related to dinitrogen oxide (Table 1).

CONCLUSION

An approach for modeling and restructuring datasets concerning pollutant emissions in the air is presented in the current paper. It integrates the processes related to searching the necessary information from xlsx files, grouping and organizing the studied datasets, as well as entering and storing them in the built database. Certain criteria are defined in order to form and extract groups of elements from one or more tables of the database, as well as to establish the relevant dependencies for the considered objects in Bulgaria for the period from 2011 to 2020. Analysis of variance and Tukey's range test are used for statistical assessment of the examined datasets.

The results for the studied data show a decreasing tendency in the values of sulphur and nitrogen oxides in the air from combustion processes. The same conclusions are drawn for the investigated emissions of methane, ammonia and non-methane volatile organic compounds in the air from waste and waste water treatment. An increase in the values of the nitrogen oxides, dinitrogen oxide and ammonia in the air from agriculture is calculated from about 63.89%, 51.25% and 17.51% respectively for the time interval 2011-2016. The reverse process is observed for the next group of elements during the time period 2016-2020. A gradual decrease in the values of the carbon oxide and non-methane volatile organic compounds in the air from road transport is calculated of about 37.47% and 43.14%.

The statistical evaluation of the studied information concerning pollutant emissions in the air for the considered ten years period shows groups with statistically significant differences.

The presented approach can also be applied to examine datasets on air pollutant emissions for other countries.

REFERENCES

1. R. Ioan, P. Raluca, “The Development of Information and Communication Technologies Sector in the Context of the New Economy”, *Business Management Dynamics*, (2013), vol. 3, Issue 4, pp. 1-6, ISSN: 2047-7031
2. M. Polyakov, I. Khanin, V. Bilozubenko, M. Korneyev and N. Nebaba, “Information Technologies for Developing a Company’s Knowledge Management System”, *Knowledge and Performance Management*, (2020), 4(1), pp.15-25, http://dx.doi.org/10.21511/kpm.04(1).2020.02
3. V. Sugumaran, “Intelligent Information Technologies: Concepts, Methodologies, Tools and Applications”, IGI Global, USA, (2007), ISBN:978-1-59904-941-0
4. D.-C. Arghir, “Solutions for Urban Resilience Assessment in Crisis Time”, *Proceedings of the International Conference on Business Excellence*, Sciendo, (2021), Vol. 15(1), pp. 1113-1126, doi: 10.2478/picbe-2021-0104, ISSN 2558-9652
5. D. Wackerly, W. Mendenhall, R. L. Scheaffer, “Mathematical Statistics with Applications”, 7th edition, Cengage Learning, (2014), ISBN 9781111798789
6. D. R. Anderson, D. J. Sweeney, T. A. Williams, J. D. Camm, J. J. Cochran, “Essentials of Statistics for Business and Economics”, 7th edition, Cengage Learning, USA, (2014), ISBN-13: 978-1305081598
7. K. Sumiran, “An Overview of Data Mining Techniques and Their Application in Industrial Engineering”, *Asian Journal of Applied Science and Technology*, (2018), Vol. 2, Issue 2, pp. 947-953, ISSN: 2456-883X
8. B. Ratner, “Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data”, 3rd Edition, CRC Press, Taylor & Francis Group, Chapman and Hall, USA, (2017), ISBN 9781498797603

9. A. Sharma, R. Sharma, V. Kr. Sharma, V. Shrivatava, “Application of Data Mining - A Survey Paper”, *International Journal of Computer Science and Information Technologies*, (2014), Vol. 5 (2), pp. 2023-2025, ISSN: 0975-9646
10. E. Sweetline Priya, K. Anandhan, “An Overview of Data Mining - A Survey Paper”, *International Journal of Modern Computer Science*, (2018), Vol. 6, Issue 1, pp. 19-21, ISSN: 2320-7868
11. N.S. Represa, A. Fernández-Sarría, A. Porta, J. Palomar-Vázquez, “Data Mining Paradigm in the Study of Air Quality”, *Environmental Processes*, (2020), 7, pp. 1-21, https://doi.org/10.1007/s40710-019-00407-5
12. D. Birant, "Comparison of Decision Tree Algorithms for Predicting Potential Air Pollutant Emissions with Data Mining Models", *Journal of Environmental Informatics*, (2011), Vol.17, no.1, pp. 46-53, doi:10.3808/jei.201100186, ISSN: 1726-2135
13. A. P. Yudison, R. Driejana, “Development of Indoor Air Pollution Concentration Prediction by Geospatial Analysis”, *Journal of Engineering and Technological Sciences*, (2015), Vol. 47, No. 3, pp. 306-319, doi: https://doi.org/10.5614/j.eng.technol.sci.2015.47.3.6
14. National Statistical Institute, Bulgaria, http://www.nsi.bg, Accessed on January 25th, 2023
15. E. B. Magrab, “Regression Analysis and the Analysis of Variance”, *In: Engineering Statistics*, (2022), pp. 93-124, Springer, Cham, https://doi.org/10.1007/978-3-031-05010-7_3
16. N. Keranova, “Mathematical Approaches for Study and Analysis of the Land Market in Bulgaria”, *Scientific journal «Economics and Finance»*, (2017), SAUL Publishing Ltd, Dublin, Ireland, pp. 95-99
17. J. W. Tukey, “Comparing Individual Means in the Analysis of Variance”, Biometrics, (1949), Vol. 5, No. 2, pp. 99-114, International Biometric Society, https://www.jstor.org/stable/3001913, Accessed on January 15th, 2023
18. D. Dimova, “Mathematical Forecasting Model for Employees by Regions”, *Industrial Technologies*, (2014), Vol. I (1), pp. 37-41, ISSN 1314-9911
19. D. R. Anderson, D. J. Sweeney, T. A. Williams, J. D. Camm, J. J. Cochran, M. J. Fry, J. W. Ohlmann, “Modern Business Statistics with Microsoft Excel”, 7th edition, Cengage Learning, USA, (2020)
20. J. Fox, “Using the R Commander: A Point-and-Click Interface for R”, Chapman and Hall/CRC, New York, (2016), ISBN 9781498741903